Stats 101C: Homework 1

Kitu Komya (404491375)

April 12, 2017

```
heart <- read.csv(file = "Heart.csv")
```

1a.

parameter/inference questions:

(1) Is age significant in whether a person has heart disease?
(2) Is typical chest pain significant in whether a person has heart disease?

prediction questions:

(3) Would a person with typical chest pain have heart disease?
(4) Would a person with a resting bp of 170 have heart disease?

1b.

```
heart$Sex <- as.factor(heart$Sex)
fit <- glm(AHD~ ., family = binomial(link = "logit"), data = heart)
summary(fit)
```

```
##
## Call:
## glm(formula = AHD ~ ., family = binomial(link = "logit"), data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7629  -0.5101  -0.1494   0.3460   2.7301
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.739690   2.931081  -1.617  0.10587
## X                    0.002676   0.002221   1.205  0.22811
## Age                 -0.013183   0.024785  -0.532  0.59479
## Sex1                 1.486941   0.519796   2.861  0.00423 **
## ChestPainnonanginal -1.755328   0.493018  -3.560  0.00037 ***
## ChestPainnontypical -0.951481   0.560165  -1.699  0.08940 .
## ChestPaintypical    -2.069490   0.667501  -3.100  0.00193 **
## RestBP               0.023561   0.011259   2.093  0.03638 *
## Chol                 0.005237   0.003953   1.325  0.18519
## Fbs                 -0.568851   0.597353  -0.952  0.34095
## RestECG              0.274371   0.191193   1.435  0.15127
## MaxHR               -0.019537   0.010940  -1.786  0.07411 .
## ExAng                0.746333   0.439690   1.697  0.08962 .
## Oldpeak              0.397408   0.234648   1.694  0.09033 .
## Slope                0.678110   0.372667   1.820  0.06882 .
```

```
## Ca                  1.248132   0.272769   4.576 4.74e-06 ***
## Thalnormal          0.008928   0.819267   0.011  0.99131
## Thalreversable      1.442033   0.804871   1.792  0.07319 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 193.36  on 279  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 229.36
##
## Number of Fisher Scoring iterations: 6
```

I'm going to answer question 1. I have fit a model in which the p-value for Age is 0., which means that at a significance level of 5%, Age is not significant in determining whether a person has heart disease since its p-value is greater than the significance level.

2a.

```
data <- read.csv(file = "hw1.csv")
poly1 <- lm(y~x, data = data)
poly2 <- lm(y~x + I(x^2), data = data)
poly3 <- lm(y~x + I(x^2) + I(x^3), data = data)
poly4 <- lm(y~x + I(x^2) + I(x^3) + I(x^4), data = data)
poly5 <- lm(y~x + I(x^2) + I(x^3) + I(x^4) + I(x^5), data = data)
anova(poly1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 479453  479453  38.488 0.0004436 ***
## Residuals  7  87201   12457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poly2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 479453  479453 42.0736 0.0006383 ***
## I(x^2)     1  18827   18827  1.6521 0.2460502
## Residuals  6  68374   11396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poly3)
```

```
## Analysis of Variance Table
##
```

```
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 479453  479453 39.0022 0.001542 **
## I(x^2)     1  18827   18827  1.5315 0.270827
## I(x^3)     1   6909    6909  0.5620 0.487209
## Residuals  5  61465   12293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poly4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x          1 479453  479453 104.7432 0.0005137 ***
## I(x^2)     1  18827   18827   4.1130 0.1124611
## I(x^3)     1   6909    6909   1.5093 0.2865864
## I(x^4)     1  43155   43155   9.4278 0.0372756 *
## Residuals  4  18310    4577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poly5)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 479453  479453 78.6485 0.003023 **
## I(x^2)     1  18827   18827  3.0883 0.177105
## I(x^3)     1   6909    6909  1.1333 0.365161
## I(x^4)     1  43155   43155  7.0791 0.076296 .
## I(x^5)     1     21      21  0.0035 0.956670
## Residuals  3  18288    6096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To find the MSE_training, I divided the Residuals of the Sum Squares of each anova(model) by number of observations, which is 9 for all of them.

Polynomial Order	MSE_Training
1	9689
2	7597
3	6829
4	2034
5	2032

2b.

Since the MSE_Training is the lowest for polynomial order 5, that's the model I would choose.

2c.

```
set.seed(123456)
x = seq(0, 4, by = 0.5)
y = 500 + 200*x + rnorm(length(x), 0, 100)

p1 <- predict(object = poly1, newdata = data.frame(x))
p2 <- predict(object = poly2, newdata = data.frame(x))
p3 <- predict(object = poly3, newdata = data.frame(x))
p4 <- predict(object = poly4, newdata = data.frame(x))
p5 <- predict(object = poly5, newdata = data.frame(x))

MSE <- function(y, pred_val)
{
  SSE <- 0
  for (i in 1:length(pred_val))
  {
    SSE <- SSE + (pred_val[i] - y[i])^2
  }

  MSE <- SSE/length(pred_val)
  return(MSE)
}
MSE(y, p1)
```

```
##        1
## 19129.35
```

```
MSE(y, p2)
```

```
##        1
## 21410.25
```

```
MSE(y, p3)
```

```
##        1
## 19643.81
```

```
MSE(y, p4)
```

```
##        1
## 26458.53
```

```
MSE(y, p5)
```

```
##        1
## 26385.96
```

Polynomial Order	MSE_Testing
1	19129.35
2	21410.25
3	19643.81
4	26458.53
5	26385.96

2d.

The MSE for testing data is much higher than that of the training data, which makes sense since in general, testing data can overfit while training data can underfit. Knowing the true model makes sense because the lowest MSE for the testing data is for polynomial 1 which is linear, which is the true model.

3.

(a) Flexible would be **better**. With a larger sample size, we can fit the data closer with a flexible approach since it would lower the variance by taking advantage of the larger sample size. Since the number of predictors is small, the bias will also be small.
(b) Flexible would be **worse**. With a smaller sample size, a flexible approach would overfit the data.
(c) Flexible would be **better**. Flexible models are usually non-linear to begin with, so with more degrees of freedom, it would fit the data better.
(d) Flexible would be **worse**. It would fit to the noise of the data, thereby overfitting.

4a.

(1) Should this applicant get into UCLA? Response: admit, reject, waitlist Predictors: GPA, ACT/SAT scores, essay, extracurricular activities, leadership Goal: Prediction because we are trying to know the value of our response variable.
(2) Will this couple get divorced? Reponse: yes, no Predictors: length of relationship, age, ethnicity, religion, sexuality Goal: Prediction because we are trying to know the value of our response variable.
(3) Will this student get an A in Stats 101c? Response: yes, no Predictors: attendance, grades, time studying, time spent in office hours Goal: Prediction because we are trying to know the value of our response variable.

4b.

(1) How much will a CEO make? Response: salary Predictors: industry, sex, age, time spent in industry, years of education Goal: Prediction because we are trying to know the value of our response variable.
(2) How much will a car cost? Reponse: cost Predictors: model, type, year, color, mileage, engine size Goal: Prediction because we are trying to know the value of our response variable.
(3) How many iPhone 7s will be sold in 2017? Response: count Predictors: advertisement, cost, battery life, size, weight, speed Goal: Prediction because we are trying to know the value of our response variable.

4c.

(1) What rating should this movie be given? Response: G, PG, PG-13, R Predictors: violence, sexual content, language Goal: Prediction because we are trying to know the value of our response variable.
(2) What grade should the student be given? Response: A, B, C, D, F Predictors: attendance, exam scores, time spent studying, time spent going to office hours Goal: Prediction because we are trying to know the value of our response variable.
(3) What type of music does this person listen to? Response: Pop, Hip-Hop, Rap, Classical, Jazz, R&B, Soul, Funk Predictors: age, hometown, gender, race, parent's birthplace Goal: Prediction because we are trying to know the value of our response variable.

5a.

These are the assumptions for the Gauss-Markov theorem that must be met in order to achieve a desirable quality of unbiasedness: * The parameters must be linear. * The expected value of the error term is zero for all of the observations. * There is homoskedasticity, meaning that the conditional variance of the error term is constant in all of x over time. * The error term must be independently distributed, meaning that the covariance between observations must be 0. * Xi must be deterministic, meaning that x is uncorrelated with the error term.

5b.

The Gauss-Markov theorem would be violated when the expected value of the error term is nonzero. This would lead to the intercept being biased, which would be non-ideal.